ASSOCIATION RULE MINING USING MARKET BASKET ANALYSIS

NOR FADILAH TAHAR @ YUSOFF

UNIVERSITI UTARA MALAYSIA
2003

ASSOCIATION RULE MINING USING MARKET BASKET ANALYSIS

A thesis submitted to the Graduate School in partial
fulfillment of the requirements for the degree
Master of Science (IKBS),
Universiti Utara Malaysia

by

Nor Fadilah Tahar @ Yusoff

Sekolah Siswazah
(Graduate School)
Universiti Utara Malaysia

PERAKUAN KERJA KERTAS PROJEK
(Certification of Project Paper)

Saya, yang bertandatangan, memperakukan bahawa
(I, the undersigned, certify that)

NOR FADILAH BT. TAHAR @ YUSOFF

calon untuk Ijazah
(candidate for the degree of) Sarjana Sains (Sistem Pintar Berasaskan Pengetahuan)

telah mengemukakan kertas projek yang bertajuk
(has presented his/her project paper of the following title)

ASSOCIATION RULE MINING USING MARKET BASKET ANALYSIS

seperti yang tercatat di muka surat tajuk dan kulit kertas projek
(as it appears on the title page and front cover of project paper)

bahawa kertas projek tersebut boleh diterima dari segi bentuk serta kandungan, dan meliputi bidang ilmu dengan memuaskan.
(that the project paper acceptable in form and content, and that a satisfactory knowledge of the field is covered by the project paper).

Nama Penyelia Utama
(Name of Supervisor) : En. Mohd. Shamrie Sainin

Tandatangan
(Signature) : _____ Tarikh*(Date)*: 31/07/03

Nama Penyelia Kedua
(Name of Supervisor) : Prof. Madya Fadzilah Siraj

Tandatangan
(Signature) : _____ Tarikh*(Date)*: 31/07/03

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by my supervisor(s) or, in their absence, by the Dean of the Graduate School. It is understood that any copying or publication, or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or in part should be addressed to:

Dean of Graduate School
Universiti Utara Malaysia
06010 Sintok
Kedah Darul Aman

ABSTRACT
(BAHASA MELAYU)

Perolehan pengetahuan dalam pangkalan data merupakan satu bidang yang mempunyai matlamat untuk mengekstrak pengetahuan yang berguna daripada koleksi data yang banyak. Terdapat pelbagai aplikasi yang menggunakan kaedah perolehan pengetahuan ini terutama syarikat syarikat atau aplikasi - aplikasi yang menggunakan pangkalan data yang besar dalam aktiviti seharian. Ini disebabkan oleh kaedah ini amat berguna untuk meningkatkan keberkesanan kepada syarikat terutamanya dalam strategi pemasaran. Analisis 'Market Basket ' merupakan salah satu teknik yang terdapat dalam perlombongan data yang boleh digunakan dalam strategi pemasaran. Analisis ini adalah bertujuan untuk melihat hubungan yang wujud di antara produk · produk yang dibeli oleh pengguna. Ia juga digunakan untuk mengenalpasti kelakuan pembeli dan barangan yang menarik minat pembeli. Hasil daripada analisis ini dapat membantu peniaga dalam membuat promosi kepada barangan tertentu, memperbaiki susun atur barangan di dalam kedai dan juga membantu dalam strategi pemasaran silang. Untuk merealisasikan analisis ini, perlombongan petua hubungan merupakan teknik yang popular digunakan untuk analisis market basket. Ini disebabkan oleh kebolehan yang ada pada kaedah ini dan ia boleh digunakan untuk mengekstrak petua daripada semua transaksi yang ada. Kajian ini adalah bertujuan untuk menganalisis transaksi -- transaksi menggunakan kaedah analisis market basket untuk menghasilkan petua berkaitan dengan barangan yang biasa dibeli oleh pengguna. Analisis ini juga menggunakan kaedah yang statistik. Hasil daripada kedua -- dua kaedah akan di bandingkan dan seterusnya 'rules ' yang diperolehi dengan menggunakan petua hubungan akan disahkan berdasarkan hasil yang diperolehi daripada kaedah statistik.

ABSTRACT
(English)

Knowledge discovery in databases (KDD) is afield whose goal is to extract usable knowledge from a collection of data (Pazzani et al., 1997). Many applications have applied knowledge discovery especially for the company or any applications that use large database in their company activities. This is because that knowledge discovery is very important to extract the useful knowledge in order to improve marketing strategy. Market Basket Analysis (MBA) is one of data mining techniques thut can be used in marketing strategy. Its purpose is to find interesting relationships among retail products. Market basket analysis is used to understand customers buying habits and preferences. The results of this analysis can help the retailers to design promotions, arrange shelf or catalogue items and develop cross marketing strategies. To do the analysis, association rules mining is the popular technique for market basket analysis because of their potential in extracting rules between items in transactions. This study presents the analysis using market basket to find the new rules of items that purchased together in transactions. The result of statistical analysis is also presented in this studying in order to compare and then to validate the rules that obtainedfrom apriori algorithm.

ACKNOWLEDGEMENTS

The author wishes to thank the Management of one mini market in Terengganu that allowing this study to be carried out, and the use of items in transactions as samples in this study. Thanks also to the staffs at that mini market for their cooperation in doing this study.

A special thanks to my two supervisors, Mr. Mohd Shamrie bin Sainin and Assoc. Prof. Fadzilah Siraj, for the valuable assistance in guiding me through this study.

A deepest appreciation to my family and my friends who have been supported and understanding throughout my course of study.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1
Introduction

1.1 Context Of The Study

Knowledge Discovery of Database (KDD) is the process of looking in the database to find hidden knowledge patterns without a predetermined idea or hypothesis about what the patterns may be (Chen, 2001). For example, the database stores the information transaction patterns done by the customers and the initiative of KDD is to find the interesting patterns where user does not have to address the relevant questions first.

Data Mining is the non-trivial extraction of implicit, previously unknown, interesting and potentially useful information (usually in the form of knowledge patterns or models) from data. The extracted knowledge is used to describe the hidden regularity of data, to make prediction or to aid human users in other ways. The ever-increasing popularity of data mining is due to demands from various real world applications in decision-making. The objective of data mining can be categorized into two general categories; prediction and knowledge discovery. Discovery or description focuses on finding human interpretable patterns describing the data. The relative importance of prediction and description for particular data mining can vary considerably. However, in the context of knowledge discovery, description tends to be more important than prediction. This is in contrast to pattern recognition and machine learning applications, where prediction is often the primary goal. While prediction, involves using some variables or fields in database to predict unknown or future values of other variables of interest.

To best apply the data mining techniques, it must be fully integrated with a data warehouse as well as flexible interactive business analysis tools. Many data mining tools currently operate outside of the warehouse, requiring extra steps for

The contents of the thesis is for internal user only

REFERENCES

1. Ahmad, A. M., Manaf, S. A. & Hijazi, A. M. H., (2002). An Application Of Multi Agent System For Data Mining Using Market Basket Analysis Technique, *In Proceedings Of The International Conference On Artificial Intelligence in Engineering and Technology.*

2. Aggarwal, C. C. & Yu, P. S., (1998). A New Framework For Itemset Generation. In Proceedings ACM PODS Conference.

3. Agrawal, R. & Srikant, R., (1994). Fast Algorithms for Mining Association Rules In Large Databases. *In Proceedings Of 20th International Conference on Very Large Database.*

4. Apte, C., Liu, B., Pednault, E. P. D. & Smyth, P., (2002). Business Applications Of Data Mining.
   URL:www.research.ibm.com/dar/papers/pdf/ Date Accessed: 23th July 2003

5. Berry, Michael J. A., & Linoff, G., (1997), *Data Mining Techniques For Marketing, Sales and Customer Support*, John Wiley Sons, Inc.

6. Bhutta, K. S., (2000). Data Mining: Applications In Business.
   URL: www.sbaer.uca.edu/Research/2002 Date Accessed: 27th July 2003.

7. Borgelt, C. & Kruse, R., (2002). Induction of Association Rules: Apriori Implementation. URL: http://www.citeseer.nj.nec.com
   Date Accessed: 19th May 2003.

8. Brijs, T., Swinnen, G., Vanhoof, K., & Wets, G., (1999). Using Association Rules For Product Assortment Decisions in Automated Convenience Stores, http://www.citeseer.nj.nec.com Date Accessed: 24th May 2003.

9. Brin, S., Motwani, R., & Silverstein, C., (1997). Beyond Market Baskets: Generalizing Association Rules To Correlations. *ACM SIGMOID-'97.*

10. Chen, Z., (2001), *Data Mining and Uncertain Reasoning: An Integrated Approach*, John Wiley and Sons, Canada.

11. Chen, M. S., Han, J. & Yu, P. S., (1996). Data Mining: An Overview From A Database Perspective, *IEEE Trans. On Knowledge And Data Engineering.*

12. Cryan, M. E., (1999). Learning and Approximation Algorithms for problems motivated by Evolutionary Trees.
URL: http://www.citeseer.nj.nec.com Date Accessed: 20th July 2003

13. Cunningham, S. J. & Frank, E., (1999). Market Basket Analysis of Library Circulation Data, *In Proc. Of The Sixth International Conference on Neural Information Processing,* Vol. 2, Perth, Australia, IEEE Service Center.

14. Cunningham, S., Sreedhar, S. & Smart, B., (1997). Completing A Solution For Market Basket Analysis, *Fourth International Conference On Knowledge Discovery And Data Mining.*

15. Dehaspe. L., Toivonen. H. and King. R. D. (1998), Finding Frequent Substructures in Chemical Compounds.
URL: http://www.citeseer.nj.nec.com. Date Accessed: 1 April 2003.

16. Doddi, S., Marathe, A., Ravi, S.S. & Toney, D. C., (1999). Discovery of AssociationRules in Medical Data. URL: http://www.c3.lanl.gov Date Accessed: 15th June 2003.

17. Dunkel, B., & Soparkar, N., (1999). Data Organization and Access for Efficient Data Mining. *Technical Report*, The University of Michigan, Ann Arbor.

18. Dyche. J. (2000), *e-Data: Turning Data Into Information with Data Warehouse*, Addison Wesley Longman, Second Edition, Canada.

19. Gayle, S., (2000). The Marriage of Market Basket Analysis to Predictive Modeling, URL: http://www.citeseer.com. Date Accessed: 20th May 2003.

20. Goransson, O., (2003). Market Basket Analysis. URL: http://www.megaputer.com/products/pa/algorithms/ba.php3 Date Accessed: 15th May 2003.

21. Han, J. & Kamber, M., (2001). *Data Mining: Concepts And Techniques*, Morgan Kaufmann Publishers.

22. Hao, M. C., Hsu, M., Dayal, U., Wei, S. F., Sprenger, T. & Holenstein, T. (2000), Market Basket Analysis Visualization on a Spherical Surface. URL: http://www.hpl.hp.com/techreports/2001/HPL-2001-3.pdf Date Accessed: 1st April 2003.

23. Hipp, J., Guntzer, U. & Nakhaeizadeh, G., (2002). Data Mining Of Association Rules And The Process Of Knowledge Discovery in Database, *ACM SIGKDD Explorations.*

24. Joshi, M. V., Karypis, G. & Kumar, V., (1997). Parallel Algorithms for Mining Sequential Associations: Issues and Challenges. URL: http://www.citeseer.nj.nec.com Date Accessed: 20th June 2003.

25. Lee, K. B. & Suh, S. C., (1998). The Efficient Algorithm in The Volume of Market Basket Data For Association Rules, *Expersys-98, The 10th International Conference.*

26. Lee, K. B. & Suh, S. C., (1999). Integration Of A Regression Analysis With Association Rules For Effective Data Mining, *Expersys-98, The 10th International Conference.*

27. Lin, W., Alvarez, S. A. & Ruiz, C., (2001). Collaborative Recommendation via Adaptive Association Rule Mining. URL: http://nas.cl.uh.edu/boetticher/ML_DataMining/alvarez.pdf Date Accessed: 24th July 2003.

28. Little, B. B., Johnston, W. L., Lovell, A. C., Rejesus, R. M. & Steed, S. A., (2001). Collusion in The U. S. Crop Insurance Program: Applied Data Mining, http://www.cse.unsw.edu.au. Date Accessed: 20th June 2003.

29. Lu, W., (2000). Best Buy's. URL: https://www.sun.com/products-n-solutions Date Accessed: 28th July 2003.

30. Maclean, I., (2003) The Use Of Statistics In Business. URL: http://www.statisticsforbusiness.co.uk/ essential/edsbusch1.pdf Date Accessed: 24 June 2003.

31. Mannila, H., (1997). Methods And Problems In Data Mining, *In Proceedings Of The International Conference On Database Theory*, Springer-Verlag.

32. Michail, A., (1999). Data Mining Library Reuse Patterns In User Applications. URL: http://www.citeseer.nj.nec.com Date Accessed: 30th April 2003.

33. Mitra, S., Pal, S. K., & Mitra, P.,(2002). Data Mining in Soft Computing Framework: A Survey, *IEEE Trans. Neural Networks*, Vol. 13.

34. Montes-y-Gomez, M., Gelbukh, A. & Lopez-Lopez, A., (2001). Mining The News: Trends, Associations and Deviations. *17th International Joint Conference On Artificial Intelligence, IJCAI-01, Workshop On Adaptive Text Mining.*

35. Moore, D. S., (1986). Tests Of Chi-Squared Type. In: R. B. D' Agostino and M. A. Stephens (eds), *Goodness of Fit Techniques*, Marcel Dekker, New York.

36. Park, J. S., Chen, M. S. & Philip S. Y., (1995). An Effective Hash Based Algorithm For Mining Association Rules. *In Proceedings Of The ACM SIGMOID International Conference on Management of Data.*

37. Pazzani, M., Mani, S. & Shankle, W. R., (1997). Comprehensible Knowledge Discovery in Database. *Cognitive Science Conference.*

38. Rantzau, R., (1997). Extended Concepts For Association Rule Discovery. URL: http://www.citeseer.nj.nec.com Date Accessed: 26th July 2003.

39. Redlon, M. F., (2000). Market Basket Analysis Macro: The "Poor Man's Recommendation Engine".
URL: www.lex-jansen.demon.nl/sugi/html/aqr.html.
Date Accessed: 25th July 2003.

40. Silverstein, C., Brin, S., Motwani, R. & Ullman, J., (1998). Scalable Techniques For Mining Causal Structures.
URL: www-cs-students.stanford.edu/~csilvers/papers/causality-dmkd.ps
Date Accessed: 27th June 2003.

41. Shi, A., Long. A. & Newcomb, D., (2001). Enhancing e-Business Through Web Data Mining,
URL: **http://eric.univ-lyon2.fr/~pkdd2000/Download/WS1_12.pdf**
Date Accessed: 23th July 2003.

42. Srikant, R., Vu, Q. & Agrawal, R., (1997). Mining Association Rules with Item Constraints, *In Proceedings 3rd International Conference Knowledge Discovery And Data Mining.*

43. Thearling, K., (2000). An Introduction To Data Mining: Discovering Hidden Value In Your Data Warehouse.
URL: **http://www.thearling.com/text/dmwhite/dmwhite.htm**
Date Accessed: 25th July 2003.

44. Webb, G. I., (2000). Efficient Search For Association Rules. *To Appear In Proc. KDD – 2000, Boston.*

45. Srikant, R. & Agrawal, R., (1996). Mining Quantitative Association Rules In Large Relational Tables. *In Proceedings Of The ACM SIGMOID Conference On Management Of Data.*

46. Wuthrich, B., (1995). Knowledge Discovery in Databases. *Technical Report, The Hong Kong University of Science & Technology.*

47. Yeo, D. (2000). Better Insuring With Information, *Canadian Underwriter.*

48. Zhu. H. (1995), Online Analytical Mining of Association Rules.
URL: http://www.citeseer.nj.nec.com. Date Accessed: 1 April 2003.